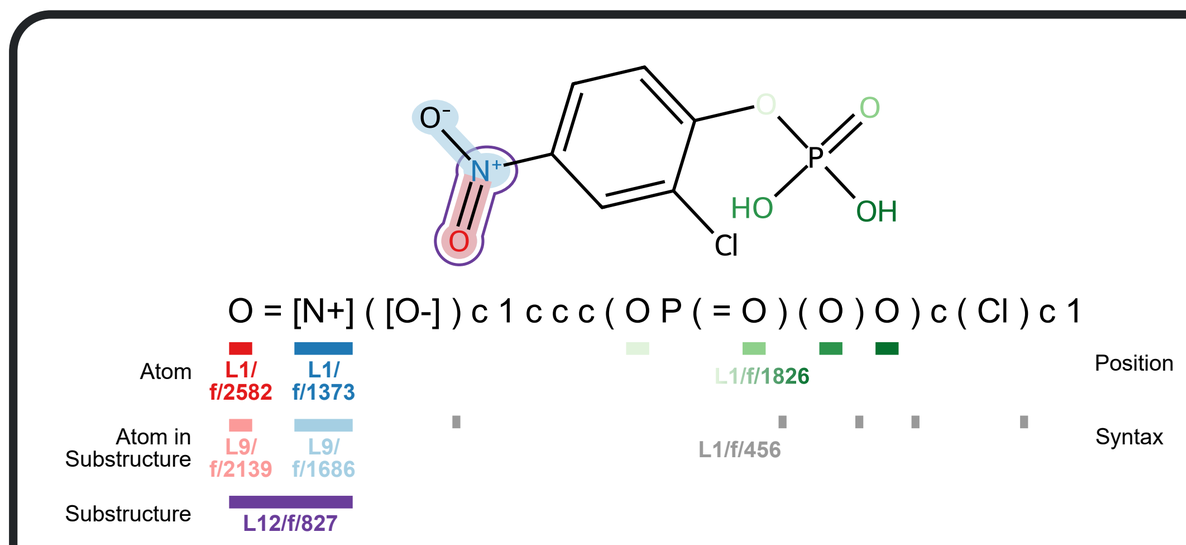


MolFormer builds molecules bottom-up, from atoms to atoms-in-substructure to substructures, analogous to molecular fingerprints. Position latents carry SMILES syntactic information from early layers, shaping final-layer representations.

What Does a Chemical Language Model Know About Molecules?

Christian Kenneth[1], Etowah Adams[2], Liam Bai[3], Gerard JP van Westen[4]

[1]Independent, [2]Columbia University, [3]Generate:Biomedicines, [4]Leiden University

Preprint

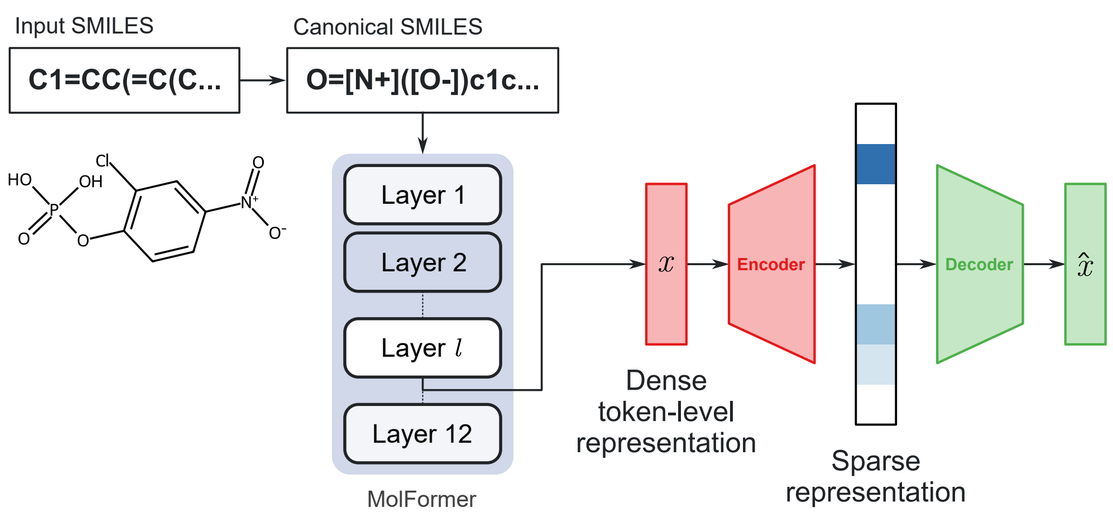
Visualizer

Looking Inside the Black Box

Do chemical language models (cLMs) learn molecular features, or just SMILES string patterns? Yet good benchmark scores alone cannot answer this.

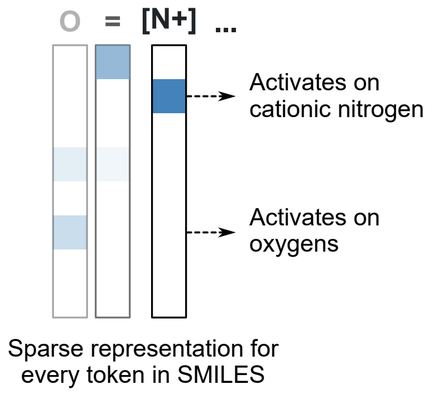
Answering this requires examining what cLMs represent internally, not just what they predict. Mechanistic interpretability has done this for protein (ESM2) and nucleic acid (Evo2) language models, but **cLMs remain unexplored**.

Methods

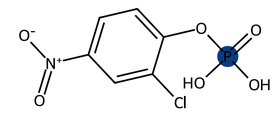


► k-SAE (k = 128; d = 3,072) trained on MolFormer residual streams at layers 1, 3, 6, 9, and 12, using ~2.2 million SMILES.

Structure and Syntax Concepts



Sparse representation for every token in SMILES

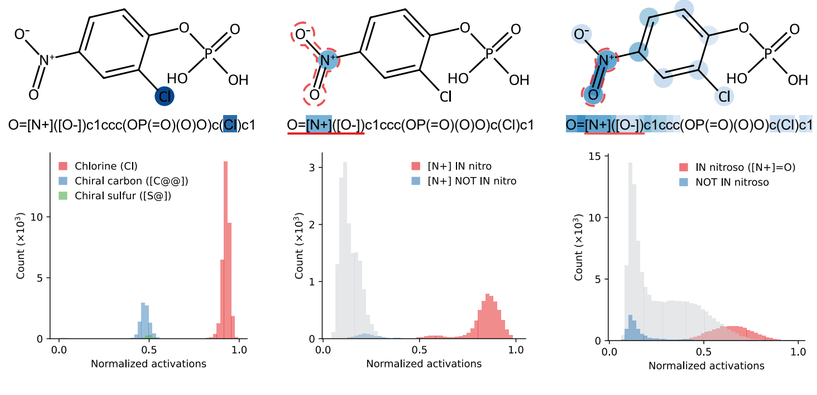


Molecular concepts discovery

...(OP(=O)(O)O)c(Cl)c1 Bond
...(OP(=O)(O)O)c(Cl)c1 Branch
...(OP(=O)(O)O)c(Cl)c1 Ring
...(OP(=O)(O)O)c(Cl)c1 Position

Syntax parsing discovery

► We evaluate molecular concepts using F1 score, and syntax features using token Simpson's dominance, Spearman correlation, and token exclusivity proportions.



Position Latents
L6/f/963; $\bar{\rho} = -0.99$
Branch closings
...(OP(=O)(O)O)c(Cl)c1

Anything-But-X Latents
L1/f/2597
Anything but aromatic carbons, rings, and branch closings
...(OP(=O)(O)O)c(Cl)c1

L3/f/172
Anything but nitrogens, oxygens, and branch tokens
...(OP(=O)(O)O)c(Cl)c1

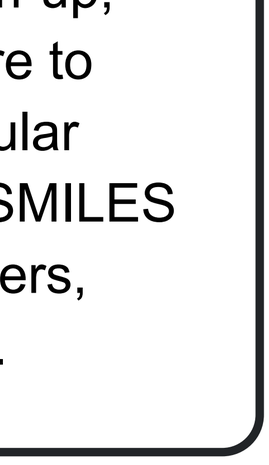
Same Molecule, Different Encoding

O=[N+]([O-])c1ccc(OP(=O)(O)O)c(Cl)c1 RDKit Canon

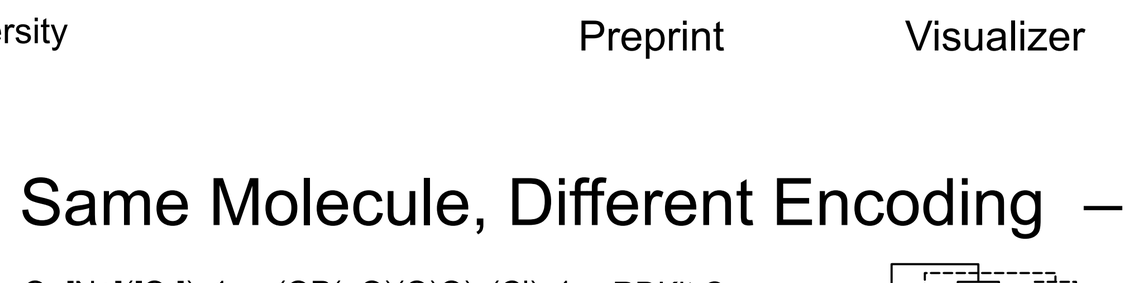
Augmented SMILES
O(P(O)(=O)O)c1ccc(cc1Cl)[N+]([O-])=O Non-canon
Clc1cc([N+]([O-])=O)ccc1OP(O)(=O)O Non-canon (Short)
c1c([N+](=O)[O-])cc(c(c1)OP(=O)(O)O)Cl Non-canon (Long)

► We measure each latent's effect size as the standardized mean difference (SMD) between canonical and non-canonical SMILES.

**Results for corrupted SMILES are omitted, as non-canonical SMILES produce a larger shift.*

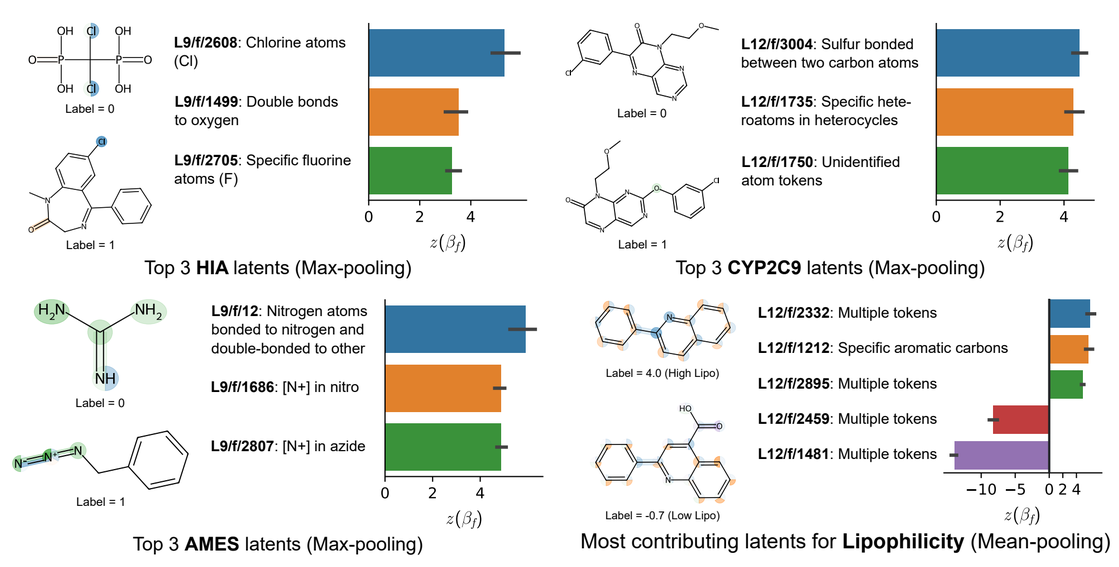


Position latents are early-layer features with peak impact on non-canonical SMILES representations in later layers.



Probing Drug-Related Properties

Pooled SAE representations outperform dense MolFormer, molecular fingerprints, and descriptors on 16 of 21 drug-related tasks.



Discussion and Limitations

► Molecular representations are built hierarchically — atom to substructure — within both structural and syntactic context.
► SMILES augmentation and multimodal representation learning may mitigate representation variance introduced by position latents.

► Analysis is limited to the residual stream; RoPE hinders attention-level circuit tracing.
► Results are preliminary as we are extending this study to ChemBERTa to generalize across encoder-only cLMs.

Mechanistic Interpretability Workshop, ICML 2026

kennethasikinnn@gmail.com